

Apparatus and Method for Disambiguating a Plurality of Entity Mentions in a Document

Field of Invention

5 The invention relates to an apparatus and a method for disambiguating a plurality of mentions identified in a document to, for instance, obtain entities and relations for storing in a database. More particularly, the invention may be directed to a computer apparatus or system and a computer-implemented method for creating and maintaining disambiguated databases of entity mentions and mention relations obtained from one or more documents.

Background

10 Entity mention disambiguation is a process of resolving which unique entities (e.g., persons, organizations or places) are the intended subjects of certain mentions in a document (e.g. Annual Report, agreement, contract, newspaper article) related to a certain entity and its related entities. A word or phrase, on its own, can be ambiguous because it may be associated with a plurality of different subjects. For instance, consider the difference between the "S R Nathan School of Human Development", and "S R Nathan, School of Human Development", where the earlier is an institution and the latter is a professor at a university department.

20 In another instance, entity mention disambiguation is difficult when a single entity, such as a person, organization or place, is routinely identified by or associated with a multitude of different words, phrases and/or abbreviations. For example, companies and organizations often have multiple trade names, abbreviations, nicknames or acronyms, while some company names are frequently misspelled. A good example will be International Business Machines Corporation, or otherwise known as "IBM", whom had previously traded as the Computing-Tabulating-Recording Company (CTR).

30 Furthermore, it would be difficult to perform entity resolution (or determine who is who) when a large number of people share the same name, when a famous individual (e.g., Mr. Michael Jackson) shares a name with non-famous individuals, and when a single individual (e.g. "Donald Trump") is associated with potentially many different organizations simultaneously or consecutively over time.

While a human reader would be able to resolve an ambiguous entity by using the context in which the ambiguous mentions appear, it is challenging to build or maintain an automated system that is reliable in disambiguating entity mentions in a lengthy document.

Summary

35 The invention is defined in the independent claims. Some optional features of the invention are defined in the dependent claims.

Brief Description of the Drawings

Examples in the present disclosure will be better understood and readily apparent to one skilled in the art from the following written description, by way of example only and in conjunction with the drawings, in which:

5

Figure 1 illustrates a flow chart of an entity mention disambiguation process according to one example of the present disclosure.

Figure 2 illustrates a schematic view of how entity disambiguation is carried out according to one example of the present disclosure.

10 Figure 3A illustrates an example of a relationship map generated during the entity mention disambiguation process according to one example of the present disclosure.

Figure 3B illustrates an example of a relationship map generated (including a large touch-point entity) during the entity mention disambiguation process according to one example of the present disclosure.

15 Figure 4 illustrates an example of a first entity relationship map, a second entity relationship map and a third entity relationship map generated during the entity mention disambiguation process according to one example of the present disclosure.

Figure 5 illustrates an example of a scoring mechanism used to identify a list of suitable entity matches for entity mention disambiguation process according to one example of the present disclosure.

20 Figure 6 is an extract of a page from a sample document analyzed by the entity mention disambiguation process according to one example of the present disclosure.

Figure 7 is a schematic view of a server or server machine that is configured to run services associated with the software components of a system to extract entity relationships ("SEER") or a user apparatus that is communicatively coupled to the server according to one example of the present disclosure.

25 Detailed Description

In an example of the present disclosure, there is provided a system to extract entity relationships ("SEER"). The system comprises a number of interconnected software components, with the flow of data between them orchestrated by a central component that is configured to extract structured data about people, companies and events from unstructured text input. The system may comprise more than
30 one hardware components, for example, more than one computer server cooperating to extract entity relationships. In another example, there may be provided an apparatus, for example, one server computer, instead of a system to perform the extraction of entity relationships. The apparatus may be a component in a system. In the following description, the example involving the SEER will be covered.

35 The purpose of the system is to extract a relationship map between all the people and companies (the "entities") mentioned in a document, along with the types of relationships that connect them. This task is of major importance in businesses such as banking, where "Know Your Customer" ("KYC") rules require an understanding of how a potential customer is connected to others.

The objective of the disclosed system is to improve accuracy in the disambiguation of entities, wherein each entity mention is a text string obtained from an input document (e.g. a company annual report). If the entity disambiguation conditions are satisfied, a database is updated with accurate entity information and entity relationship after the disambiguation of each entity mention. The database comprises verified entity relationship data indicating validated links (or relationships) between different validated entities. Unless otherwise stated, mentions of “the database” would refer to this database. The database may be stored in a local server or a remote server, and the server may comprise a plurality of computing devices communicatively coupled to one another. The disclosed SEER is configured to make use of the differences in the statistical properties of entity names (e.g. some names are very rare, while some are very common) and the connectedness of the verified entities within the database. By using several stages of disambiguation with different tradeoffs of precision and recall, the performance of entity disambiguation process improves.

In one example, the system processes PDF or Microsoft Word documents (such as Annual Reports) and derives relationship maps between accurately disambiguated entities that can be combined into a centralized Relationship Map that is of significant commercial value. Figure 2 illustrates a workflow of the SEER or system 200. With reference to Figure 2, a text document 202 (e.g. a file containing text) is inputted into the system 200 to obtain entity mentions, for example, Entity A (204a) and Entity B (204b). A relationship 206 between Entity A and Entity B is then determined from the text document 202. The unresolved (ambiguous) entity mentions 204a and 204b are matched against the database (also known as Central Index) in a step 208. If there is exact match between each unresolved entity mention and a verified entity in the database, the unresolved entity mentions becomes disambiguated. In this example, the entity mention 204a is disambiguated as disambiguated entity 210a and the entity mention 204b is disambiguated as disambiguated entity 210b. A single disambiguated relationship 212 is generated between the disambiguated entity 210a and the disambiguated entity 210b using the determined relationship 206. The system 200 is configured to disambiguate each entity mention in the text document 202 until all entity mentions are reliably disambiguated. Thereafter, the plurality of single disambiguated relationships 212 are then collated into a centralised relationship map 214.

In the present disclosure, the term “entity mention” or “mention” is a string of text that can be found in a source document that one or more algorithms, such as Named-entity recognition NER algorithm(s), have identified as being likely to be an entity. Examples such as "Sellapan Ramanathan", and "S R Nathan" could be identified from 'PERSON' NERs, and are all distinct Mentions.

In the present disclosure, the term “disambiguated entity” or “resolved entity” refers to a specific actual person or company in the database that comprises verified entity relationship data indicating links between different entities. Notably, the entity relationship data has already been verified to be correct or true by machine or human. Each disambiguated entity is associated with a unique identifier in the database. The unique identifier may be a pointer to and/or may include, for instance, data related to the disambiguated entity (such as date-of-birth, or company registration number), potential name aliases,

and/or connection information (such as "attended-university-at", or "is-married-to", which would link a person to an institution, or another person respectively). For instance, the mentions "Sellapan Ramanathan", and "S R Nathan" can point to a single Disambiguated Entity if his wife "Urmila Nandey" and/or his past employment as President of Singapore was mentioned in the document.

In the present disclosure, the term "list of entity mentions" is a list of mentions identified by algorithms such as NER(s) across the entire document. At any stage in the process of entity disambiguation, a list of currently disambiguated entities and a list of currently-eligible entities are maintained. A list of currently disambiguated entities refers to a list of mentions and their corresponding disambiguated entities, while a list of currently-eligible entities refers to a list of mentions that have yet to be disambiguated (i.e. unresolved mentions). With reference to the SEER or system 200 of Figure 2, it is configured to disambiguate each unresolved mention, based on the evidence in the documents, the verified entity relationship data in the database, and the list of currently disambiguated entities (which are assumed to be 'known to be true' at each stage in the process).

A flow chart of data flow 100 between the interconnected software components of the system 200 of Figure 2 is illustrated in Figure 1. With reference to Figure 1, while analyzing the entities mentioned in the document and the relationships between the entities (for the purpose of disambiguation), the system iteratively refines its answers (i.e. refines the candidates for disambiguating the unresolved mentions).

The functionality of these modules is described in the following paragraphs.

During an initialization stage, the system or SEER (200 of Figure 2) performs a self-check to ensure that all necessary software services are online and operable. At a step S102, a document is selected for processing. The document may be in a portable document format (PDF), an OCR-readable ("OCR" refers to "Optical Character Recognition") format, or the like.

Text extraction is performed on the selected document at a step S104, where the document is parsed to obtain sentences contained in the document as completely as possible. In one example, each page of the input document is converted into a series of textual strings. This conversion can be done either using the character-based text information in the PDF document, or via OCR. The end result obtained for each page of the document from the step S104 is a list of words, word location information relative to a bounding box or boxes (if the text conversion technique involves such box or boxes), and font information. For example, each word or a plurality of consecutive words may have location information relative to respective left, right, top and bottom bounding boxes. This allows the following steps to make use of the typographical layout of the PDF document.

The converted words are then combined into larger groupings (while retaining the structure of the text within the document) for processing. This is done by using heuristics that capture hierarchical features of the text, such as headers, subjects and actual body of the text; columns of text; tables; and bullet-pointed lists.

After step S104, a series of text strings can be outputted in a hierarchical data structure, which preserves logical relationships between elements. For instance, the logical relationships may comprise one or more of the following:

- 5 • the document is a list of pages
- each page contains a series of columns or areas, for which 'reading order flow' information is estimated
- each area contains a number of headings (as its 'children')
- each heading may have 'child' subheadings
- 10 • a paragraph may be a 'child' of a superior heading
- a table is a matrix of cells containing text
- a bullet-point list is a list of child text elements

15 Subsequently, tokenisation occurs at a step S106. At this step, the series of text strings are converted into a more uniform format. In other words, the tokenisation process ensures all textual data within the system (which includes the textual material used for machine learning model training and the extracted text) are pre-processed in an identical manner so that subsequent processing stages can be executed smoothly. For example, special characters are combined in the following:

- different bullet point types may be converted into only one bullet point type
- 20 • printer's ligature combinations are split into separate letters (such as 'fi' → 'f' and 'i').
- Quotation marks are converted into a uniform style

25 Columns of text are combined into paragraphs by removing unnecessary line breaks, and recombining hyphenations introduced by columnar formatting. Paragraphs are then split into sentences based on both punctuation and their layout on the page. For instance, "A hamburger costs \$5.00 in the U.S.A. But is more expensive in Singapore." would be correctly split into "A hamburger costs \$5.00 in the U.S.A." and "But is more expensive in Singapore."

30 At a step S108, one or more algorithms such as the following listed Named Entity Recognisers ("NERs" or NER matching algorithms) are used to identify a plurality of potential entity mentions within the parsed text obtained after S104:

- i. LegalDef – which finds legally defined terms and their aliases from phrases such as 'ABC Corporation (the "Company")'
- 35 ii. Signpost – which finds names due to specific words that are closely associated with specific entity types. For instance, 'Mr.' frequently precedes a person's name, and 'Corporation' frequently follows an organisation name.

iii. MITIE NER – An Open Source NER utility that can be used to identify which words in an input sentence correspond to an entity name, and predict the entity type from the words' context.

iv. spaCy NER – An Open Source NER utility that can be used to identify which words in an input sentence correspond to an entity name, and predict the entity type from the words' context.

It should be appreciated the above list of NERs is not exhaustive and any NER can be used to identify and obtain specific entity mentions from the parsed text. In one example, an entity mention that appear across the different NERs used may be ranked higher than an entity mention that is found by one of the NERs.

1) Kick-off Matching

Initial entities are extracted at a step S110 by identifying entities mentions from the parsed text that have an exact match against an existing entry within the verified entity relationship data in the database. An exact match occurs when a mention matches an existing entry of the database exactly character-wise and that it matches only with a single entity (which is associated to a corresponding Globally Unique Identifier (GUID) key created for the single entity).

The operation of the step S110 of the process may be as follows:

For each string identified as an entity mention by one or more of the NER matching algorithms:

(a) Find an exact character-by-character match within the verified entity relationship data:

(i) Either for the legal name of the entity;

(ii) Or one of the confirmed aliases of the entity.

(b) If no match is found, move on to the next string, leaving the current one unresolved.

Step S110 is deliberately stringent (high precision and low recall). Preferably, the name precision of the verified entity relationship data in the database should be kept very high to ensure high precision. For instance, if it is known that there are many entities containing the terms "Hour Glass", the database should be updated such that none of them would be recorded as having the generic alias of "Hour Glass" as a confirmed alias. Otherwise, the execution of step S110 for a document containing mentions of many entities containing the terms "Hour Glass" may result in imprecise match strings. An example of an extracted page in such a document is shown in Figure 6. In Figure 6, there are several mentions of different subsidiaries of the same company with the common prefix "The Hour Glass". All the entities containing the terms "Hour Glass" shown in Figure 6 are distinct and are not substitutes for one another. While it would be a common practice for the state of the art techniques to include generic terms (like 'Google' or 'Apple') as a matchable entity string alias for machine reading of, for instance, newspaper articles, this is not a usable level of precision for step S110. In another example, the system may be configured such that if many entities containing the same terms are identified in the parsed text, then

step S110 will not rely on a confirmed alias of an entity in the database containing the same terms to determine exact match. That is, (a)(ii) above will not be applied and only (a)(i) is applied.

Advantageously, in step S110, the disclosed system is able to identify a 'kick-off' set of entities, where the unique exact matches are assumed to be validly disambiguated entities. This initial determination allows the system to anchor the subsequent entity and relationship extraction phases against verified entries in the database and is found to be beneficial to overall system performance. If mistakes are made at this stage, they may be recovered at later stages. For example, two main types of fixes/recovery are available: (i) if the system finds that a 'kick-off' entity has no discoverable relationships later, the entity will be discarded as a valid disambiguated entity; and (ii) the system (e.g. 200 in Figure 2) or a human updates the database later to prevent spurious matches and the entity wrongly assumed to be a valid disambiguated entity based on such spurious matches gets discarded. For example, "MOM", which is a known alias for Singapore's Ministry of Manpower in Singapore, may be too loose an alias for entity validation globally. Such known alias recorded as a confirmed alias in the database may be deleted as a confirmed alias so that it will not be used in kick-off matching purposes under (a)(ii) above, or the system may be configured to specifically avoid use of the confirmed alias "MOM" for kick-off matching purposes i.e. (a)(ii) above will not be applied and only (a)(i) is applied for potential entities containing the term "MOM".

The system is configured to maintain a list of currently disambiguated entities and a list of currently-eligible entities. A list of currently disambiguated entities refers to a list of mentions and their corresponding disambiguated entities, while a list of currently-eligible entities refers to a list of mentions that have yet to be disambiguated (i.e. unresolved mentions). The system is configured to disambiguate each unresolved mention, based on evidence in the input document or documents, the verified entity relationship data in the database, and the list of currently disambiguated entities (which are assumed to be 'known to be true' at each stage in the process). For the example of Figure 1, if there are unique exact matches at S110, the list of currently disambiguated entities will consist of the unique exact matches. Each disambiguated entity is a single node in the verified entity relationship data.

2) Hub Entity

The following discussion of the subsequent processing steps in the flow chart of Figure 1 is based on the assumption that there is at least one exact match returned at the step S110. The objective of the subsequent processing steps is to determine a hub entity. However, it should be appreciated that if the processing at step S110 does not result in unique exact matches, the following steps can still be performed by using any entities that are disambiguated at this point (e.g. an entity mention in the parsed text of the document may match a few verified entities in the database). In another example, the system may be configured to make a series of trial assumptions that each high-frequency entity mention in the parsed text of the document is the hub entity in turn, and see how

a first entity relationship map generated by the system would perform in terms of entities confidently found (i.e. Hub Entity hypothesis). The terms “Hub entity” and “first entity relationship map” would be elaborated in the paragraphs below.

5 Determination of a hub entity is being performed at step S112 of Figure 1. One of the currently disambiguated entities (i.e. matched entity mentions in S110) is chosen as the “Hub Entity”. In other words, the hub entity is determined, from a plurality of matched entity mentions (obtained from S110), by calculating a score for each matched entity mention based on number and types of links the matched entity mention has with other matched entity mentions in the database and identifying the matched entity
10 mention with highest score as a hub entity.

The choice of hub entity depends on the following information (a) to (c):

(a) Whether the matched entity mention (obtained from S110) is a Corporation entity.

15 This can be determined based upon the classification of the NER mention returned by the NER matching algorithm(s), and/or the type of the corresponding verified entry (disambiguated entity) in the database. The classification of the NER may include PERSON, ORGANISATION, LOCATION, and DATETIME.

In one example, the SEER is configured to disambiguate entity mentions (obtained from S110) that are
20 categorized under PERSON and ORGANISATION NER types, while the other NER types (such as LOCATION) are useful for rejecting non-relevant entity mentions. For instance, "Anson Road" is a location, whereas "Anson Road Limited" is a registered United Kingdom (UK) company.

(b) A weighted sum of weighted values of one or more entity mentions that are connected by a first
25 predetermined number of hops to the matched entity mention.

(c) Whether the matched entity mention is connected to too many non-disambiguated entity mentions.

30 This is to avoid selecting a hyper-connected corporation, such as a bank or an auditor organization, as a hub entity, unless ‘proven’ more stringently in (b) above. For instance, if a matched entity mention is connected to more than a predetermined number (e.g. 30) of entities that do not have a one-hop link to other matched entity mentions, said matched entity mention cannot be chosen as a possible hub entity candidate. In other words, the weighted sum obtained for a potential hub entity would be much higher
35 than other entities that also qualify as hub entities (e.g. to allow for the document being about a hyper-connected corporation itself).

The selection of hub entity by the system (200 of Figure 2) would now be elaborated with an example illustrated by Figure 4. In this example, the system has obtained three unique exact matches (or initial
40 entities) 402, 404, and 406 at step S110 of Figure 1. These initial entities 402, 404, and 406 are

classified as Corporation entities by the system. Upon determination of the initial entities to be Corporation entities by the system, the system is configured to determine, for each matched entity mention, one or more entities that are indicated as a first predetermined number of hops away from the corresponding matched entity in the verified entity relationship data. In this example, the first
5 predetermined number is set as one for the sake of simplicity. This means that in this example, the system is configured to determine, for each of the initial entities 402, 404, and 406, one or more entities that are one hop away from the corresponding matched entity in the verified entity relationship data.

In another example, the system can be configured such that a person entity is disqualified as a hub
10 entity candidate, so as to have the system focus on disambiguating corporation entities. In this case, if an initial entity obtained from S110 is not a corporation entity but a person entity, the system does not go on to perform steps like to determine one or more entities that are one hop away from such person entity in the verified entity relationship data. The system will disqualify the person entity as a hub entity candidate. Conversely, in yet another example, if the system is configured to focus on disambiguating
15 person entities, corporation entities can be disqualified as hub entity candidates and only person entities may qualify.

In the present disclosure, "a hop" is an edge, a link or a connection in the verified entity relationship data that represents known relationships, such as "is-married-to" or "is-shareholder-of", "graduates-
20 from", "is-son-of", "is-daughter-of", "is-member-of", "is-a-subsidiary-of", "is-a-parent-entity-of", and the like. This implies that a 'one-step relationship' or 'one-hop away' from a given node/entity is the list of all nodes that are a single edge traversal from that given node. Likewise, if the system is to consider a set of nodes in the graph, then the 'one-step' nodes from that set are those nodes that are connected by a single edge traversal from any node in that original set.

Turning back to Figure 4, entities 404, 406, 412 and 414 are one hop away from the initial entity 402, entities 402 and 412 are one hop away from the initial entity 404, while entities 402, 416 and 418 are one hop away from the initial entity 406. These one hop links or connections are obtained by making
30 reference to the verified entity relationship data in the database.

A weight value is obtained for each of the one or more entities that are one-hop away from a matched entity, wherein the weight value was predetermined based on type of entity ("entity type"). For example, in one implementation, the weight value is predetermined based on the types of entities as follows:

- i. Employee of an entity;
- 35 ii. Executive director of the entity;
- iii. Shareholder of the entity;
- iv. Subsidiary entity of the entity; and
- v. Optionally, Parent entity or holding entity of the entity,

where the Executive director has the greatest weight value, the Subsidiary entity has the least weight value, and the Employee and Shareholder have the same weight value or weight values between the greatest weight value and the least weight value.

- 5 In the case that parent entity or holding entity of the entity is considered, a negative weight value may be assigned to it, since based on observation, in some situations, the Hub Entity is likely to be the parent entity or holding entity itself.

10 If there are more than one entities determined for a matched entity mention, the system is configured to calculate for the matched entity mention, a weighted sum of the weight values of the more than one entities connected via one-step relationship from a matched entity. Consequently, the weighted sum for the matched entity (initial entity) 402 in Figure 4 is the sum of w_2 , w_3 , w_4 and w_5 . Similarly, the weighted sum for matched entity (initial entity) 404 in Figure 4 is the sum of w_1 and w_4 , and the weighted sum for matched entity (initial entity) 406 in Figure 4 is the sum of w_1 , w_6 and w_7 .

15 In one example, each Employee or a shareholder connected to the matched entity (initial entity) adds 2 to the calculation of the weighted sum. Each Executive Director connected to the matched entity (initial entity) adds 3 to the calculation of the weighted sum, and each Subsidiary connected to the matched entity (initial entity) adds 1 to the calculation of the weighted sum. Additionally, if parent entity or holding entity is considered, it may subtract 1 or 2 from the calculation of the weighted sum.

20 Assuming that entity 404 and entity 406 are subsidiary entities connected to the matched entity 402, entity 412 is an Executive Director connected to the matched entity 402, and Entity 414 is a shareholder connected to the matched entity 402, the weighted sum for the matched entity 402 would be
25 $1+1+3+2=7$.

If assuming that the entity 402 is a parent entity that for example, subtracts 1 from the calculation and it is connected to the matched entity 404, and entity 412 is a shareholder connected to the matched entity 404, the weighted sum for the matched entity 404 would be $-1+2=1$. Likewise, if assuming that
30 entity 402 is a parent entity that for example, subtracts 1 from the calculation connected to the matched entity 406, entity 416 is an Executive Director connected to the matched entity 406, and entity 418 is an employee connected to the matched entity 406, the weighted sum of the match entity 406 would be $-1+3+2=4$.

35 If there are more than one matched entity mentions, the system is configured to identify the matched entity mention(s) that has the calculated weighted sum satisfying a first threshold as a hub entity candidate. In the above example, if the first threshold is 3, the matched entity 404 having a weighted sum of 1 would be filtered away. In other words, the matched entity 404 is not a hub entity candidate, and the possible hub entity candidates are the matched entity 402 and the matched entity 406, which
40 have weighted sums above 3.

After determining the hub entity candidates, the system is configured to determine a score for each hub entity candidate by counting number of links between said hub entity candidate and each matched entity mention found in the verified entity relationship data, wherein the links to be counted are the links with
5 a second predetermined number of hops between said hub entity candidate and each matched entity mention. Turning back to the example in Figure 4, if the second predetermined number of hops is set as one, the hub entity candidate 402 has two one-step links to the matched entity mentions 404 and 406 and thus has a score of 2. The hub entity candidate 406 has only one one-step link to the matched entity mention 402 and thus has a score of 1. Links are not counted for the matched entity mention 404
10 because it is not a hub entity candidate. In this case, the hub entity candidate 402 with the higher score of 2 is the hub entity.

In another example, the system may be configured to determine a score for each hub entity candidate by counting the number of entities connected to each hub entity candidate within a second
15 predetermined number of hops by making reference to the verified entity relationship data. If the second predetermined number of hops is set as two, the hub entity candidate 402 will be connected to 7 existing entries within the verified entity relationship data in the database and thus has a score of 7. The hub entity candidate 406 will be connected to 6 existing entries within the verified entity relationship data in the database and thus has a score of 6. In this case, the hub entity candidate 402 with the higher score
20 of 7 is the hub entity.

In another example, the system is configured to measure whether each hub entity candidate could be considered 'central' to the current list of disambiguated entity mentions. To do this, a "Jaccard-based" expansion factor may be calculated to work out the intersection over union of how the 'one-step'
25 relationships (from the database) of each of the resolved entities covers the current list of resolved entities. This expansion factor will be described in detail later. Then, the system sorts the list into descending order of centrality score, and, if the top score is greater than a pre-defined hurdle or threshold, accept the top-scoring entity as the hub entity.

From observation, the 'top scorer' (top-scoring entity) normally dominates over the other choices (i.e. other hub entity candidates), with other choices being far lower scoring. For instance, a Company's Annual Report may include employee biographies - even if there are shared directorships amongst the Group companies, lower level employees mentioned in the Annual Report are typically linked to the Hub Entity itself.
35

In the above examples, the system is configured to identify the hub entity candidate with highest score as a hub entity. The system would determine that the matched entity mention 402 is the hub entity because its score is highest. It should be noted that the hub entity and the matched entity mentions are deemed as disambiguated entity mentions at this stage of the entity mention disambiguation process.
40

The SEER may be configured to adapt to specific situations to ensure accurate hub entity candidate selection and accurate determination of hub entity. For instance, the SEER may be configured to process the document to obtain term definitions in the document, such as:

The Hour Glass Limited (the “Company”)

Such term definition, which is commonly seen, indicates that all mentions of the “Company” in the document refers to a company named “The Hour Glass Limited”. When such term definition is present, it may be that this company should be selected as the hub entity. In some instances, it could be that in the entire document, such term definition only shows up once, and in this case, the name of the hub entity may only be mentioned once in the document.

In another example where a unique exact match or an initial entity is not obtained at the step S110, image extraction at a preferred location (such as the corners) of the document may be initiated to obtain a possible candidate of the hub entity. For instance, the name of a hub entity in Figure 6 is on the top left hand corner of the document. However, page layout may vary in the document and cause error in obtaining entity mentions from the document. Therefore, choosing a hub entity by determining the relationships of an entity mention with other entity mentions in the document is preferred or prioritized over selection of hub entity based on text extracted from preferred locations of the document. Hub entity selection with reference to the verified entity relationship map in the database is more likely to increase the performance of the entity disambiguation process.

The SEER might be applied to legal documents, where a particular relationship may only be disclosed once over a set of documents intentionally. Therefore, it is desirable that the SEER is configured to understand the full web of relationships that exists in the document. Unlike other conventional entity-disambiguation systems which are applied to news articles or blog posts that might reasonably expect a fact to be replicated across different source documents (and thus reducing the need for completeness per-document), the SEER is configured to obtain all possible entities and the relationships between these entities. For example, none of the entities listed in the example of Figure 6 should be omitted from being considered to obtain the final results.

Based on the teachings above, the Hub Entity can be determined with a very high (e.g. >98%) certainty (i.e. high precision and recall for the selection of a single Hub entity). The Hub entity forms the basis of a hub entity cloud or a first entity relationship map, which is the set of currently disambiguated entities that are directly connected (i.e. one-hop or one-step away) by a series of relationships to the hub entity. In the example discussed in Figure 4, a first entity relationship map 410 is formed by the hub entity 402 and the disambiguated entity mentions 404 and 406. Turning back to Figure 1, this first entity relationship map 410 aids in an iterative entity coalescing and disambiguation process (i.e. steps S114 and S116) of the remaining unresolved entity mentions identified by the NERs in step S108.

Entity Coalescing step S114 and Entity Disambiguation step S116 are together an iterative process used to concretely identify the entity names found within the text with their corresponding database

entries. In one example, the system is configured to generate an ordered list of potential NER 'mentions', ranked according to criteria that prioritise those that are more likely to obtain a unique disambiguated entity. For instance, a long entity mention having a text string consisting of several very distinctive words that is mentioned in multiple places in the document, and found by multiple NER methods is more likely to be worth disambiguating early in the overall process, before other candidates. The method for ranking these candidates can either be explicitly defined using rules (such as those that require use of NER with more mentions), or machine learnt from using relevant data fed to the machine. Machine Learning methods used for the automated analysis include methods such as Support Vector Machines ("SVM"); Tree-based methods; and Neural Networks. For instance, the NER matching algorithms may be trained with a set of pre-defined learning examples so that the NER matching algorithms can identify similar candidates and rank them higher than candidates that fall out of a criterion determined from the learning data. Note that a string search by a NER matching algorithm may match multiple nodes (validated entities) in the database, but there will only be one correct disambiguation in the context of the surrounding text.

In one example, the verified entity relationship data is represented by a graph structure of entity nodes connected by arcs that represent the relationships between pairs of entities. The SEER is configured such that rules are applied to accept or reject candidate(s) of an unresolved entity mention in a list of currently-eligible entities based on their proximity (via graph connections) to previously accepted candidates in the list of currently disambiguated entities. The SEER can perform automated analysis (or the entity mention disambiguation process) to make reference to the verified entity relationship data in the database, recent disambiguation results and data currently under consideration to determine whether to accept or reject a candidate of a resolved entity mention as a disambiguated entity mention.

3) One-Step Analysis in Edge-of-cloud disambiguation

In one example of the present disclosure, an edge-of-cloud disambiguation is performed by making reference to the first entity relationship map (e.g. 410 in Figure 4) and the verified entity relationship data in the database. The edge-of-cloud disambiguation is to be performed after the hub entity is identified and is particularly desirable for entity disambiguation of very close entities or entities that have very similar characteristics. For example, unlike disambiguating the person entity 'Michael Jordan', which is a famous basketball player, from articles about basketball, it is more difficult (computationally intensive) to identify a correct Mr. Tan (which is a very common family name in Singapore) from over 10,000 candidates, all of which are doing business in Singapore. With the help of the edge-of-cloud disambiguation, the SEER can identify the correct Mr. Tan. In other words, the SEER is able to deal with 'long tail' and is not configured to prefer one entity over another during the entity disambiguation process, unless there is specific evidence from the document. 'Long tail' in the context of a distribution of numbers may refer to a portion of the distribution having many occurrences far from the "head" or central part of the distribution, such as only a few correct Mr. Tan occurrences in a list of 10000 Mr. Tan candidates, many of which are more 'popular'. Each iteration of the edge-of-cloud

disambiguation tends to give high precision results (because verified entity relationship data is used to explain/determine the hub entity cloud) and the cumulative effect is high recall (because as the cloud expands, entities that are more distant from the Hub are added).

5 The edge-of-cloud disambiguation is performed by the system on the input document to conduct disambiguation of entity mentions in the input document as follows. The edge-of-cloud disambiguation is performed under the Entity Coalescing step S114 and Entity Disambiguation step S116 shown in Figure 1.

10 Firstly, the system is configured to iteratively loop through the entity mentions identified in the input document based on frequency of occurrence, and choose an unresolved entity mention that can be disambiguated to an existing entry in the database's entities (i.e. entities in the verified entity relationship data) if the following two conditions are met:

15 (a) the candidate of the unresolved entity mention is related to at least one entity currently in the hub entity cloud (or the first entity relationship map) via a single hop (or connection; single hop can also be known as one-step); and

(b) if the unresolved entity mention is successfully disambiguated, it will expand the hub entity cloud (or first entity relationship map), when added in, by an expansion factor that falls within a predetermined range.

20 Regarding condition (b), the expansion factor should not be too small i.e. adding the entity will increase the cloud size of the hub entity cloud by just a single entity, or too big i.e. adding the entity will increase the cloud size of the hub entity cloud so much that it would shift the hub entity from being in the centre of the hub entity cloud to off-centre. An example of an entity that may cause the
25 expansion factor to be too big may be an organization similar to a bank or an accounting firm that has many relationships.

Specifically, the first entity relationship map is formed by the hub entity and the matched entities in the list of currently disambiguated entities after steps S110 and S112. Thereafter, the SEER is configured
30 to determine whether there is any entity in the first entity relationship map that is categorized as large entity with number of links above a threshold (i.e. large touch-point entities). Example of large entity could be a Bank or an Accounting/Auditing firm. These large touch-point entities tend to overwhelm the selectivity of the current entity cloud. For example, an Auditing firm is likely to have many relationships (links or arcs) across many industries, and may reduce the effectiveness of the subsequent steps if
35 added to the current entity cloud.

In addition to the generation of the hub entity cloud (i.e. the first entity relationship map), the edge-of-cloud disambiguation process comprises generating a second entity relationship map, or the 'one-step' extension of the hub entity cloud. The second entity relationship map is created using the entities in the
40 first entity relationship map, and by obtaining, from the verified entity relationship data, entities that are one hop away from the entities in the first entity relationship map. For the example illustrated in Figure

4, the first entity relationship map 410 comprises the hub entity 402 and the two matched entity mentions 404 and 406, assuming none of the hub entity 402 and the two matched entity mentions 404 and 406 are large touch-point entities. The second entity relationship map that can be created for the example of Figure 4 would be the second entity relationship map 420 as shown, which comprises the entities 402, 404 and 406 in the first entity relationship map 410 and the entities 412, 414, 416 and 418 that are one-hop away from the first entity relationship map 410. Figure 4 also shows a third entity relationship map 430 comprising the entities 412, 432, 402 and 404, which are the one-hop entities of the entity 412. The third entity relationship map 430 is a graph of connected nodes, wherein the graph of connected nodes comprises a disambiguated entity candidate and one or more entities that are one hop away from said disambiguated entity candidate.

In one example, the SEER may be configured to categorise the entity mentions in the input document into several classes after the NER matching algorithms identify the entity mentions from the parsed text of the document. In this way, the SEER may rank the entity mentions that are less computationally intensive with higher rank than the entity mentions that are more computationally intensive to disambiguate. The several classes may be as follows:

1) Too difficult to disambiguate yet

There are too many options that have little to differentiate the entity mention from others, and entity mentions in this class may be re-queued to determine whether they can be resolved later;

2) Can be disambiguated (there is a clear 'best candidate')

In the situation that the entity mention is a clear 'best candidate', the entity mention can be added to the List of Currently Disambiguated Entities

3) Not found in database

This entity mention could be either a new entity or a junk mention. A new Entity is one entity that does not have an existing database entry, and these entity mentions may be set aside for further processing or human intervention. A junk mention is an identified entity that is very unlikely to be a real entity in reality. For instance, a junk mention may be an entity mention with the name "Annual General Meeting" and which is categorised as a company. Even though the entity mention 'Annual General Meeting' may have many attributes of a company entity in a Company's report, the system is configured to filter out these mentions and report such junk mentions at the end of the disambiguation process.

Next in the edge-of-cloud disambiguation process, the SEER is configured to iteratively expand the first entity relationship map by selecting the most frequently found mention strings from the parsed text of the document (e.g. highest ranking NER mention) and fuzzy match it against the verified entity relationship data of the database based on criteria such as character-wise matching score, relationships to other known entities and whether the suggested entity (candidate entity) is one hop (relation or link) away from the entity cloud (first entity relationship map). The detailed description for this is provided below.

For example, the character-wise matches determined by the SEER for the string mention (entity mention) "Mr Charles Poon", after excluding simple honorifics such as Mr/Mrs/Ms, would include:

- Charles Poon (which is an exact match)
- Poon (Charles)
- Charlie Poon
- C Poon
- 5 • Charles XYZ
- W Poon
- W XYZ (worst match)

10 The above matches are listed in an order of character-wise similarity score (from a best match to a worst match). As can be seen from the above, the above matching algorithm used with the SEER is not a simple Levenshtein distance measurement, because it is word-aware and initial-letter aware. In another example, a character-wise similarity score may be used together with a Levenshtein distance measurement or the like to rank the unresolved entity mentions.

15 Specifically, the edge-of-cloud disambiguation process involves fuzzy matching each frequently found entity mentions against each entity listed in the database to return a list of disambiguated entity mention candidates and determining top candidates with matching similarity scores from the list of disambiguated entity mention candidates that are above or equal to a first threshold. In other words, the process is configured to identify top candidates from each frequently found entity mention in the document. With reference to Figure 5, the SEER may determine that a perfect match occurs when a candidate entity 506 matches a unique existing entry of the database (e.g. happening at S110). Only eligible candidates C_1 , C_2 and C_n above a first threshold 505 would be analyzed by the SEER to determine which eligible candidate is the most suitable for the corresponding frequently found entity mention. Candidates, such as C_{n+1} and C_{n+2} , below a second threshold 510 are considered as non-eligible candidates.

The first threshold 505 is calculated by ranking all the disambiguated entity mention candidates 502 according to matching similarity scores 504 and calculating a difference in matching similarity scores for each pair of ranked candidates consecutively, according to a formula:

30
$$D_n = C_n - C_{n+1} ,$$

wherein

D_n = difference between candidate with nth rank and candidate with one rank below the nth rank,

35 C_n = candidate with nth rank,

C_{n+1} = candidate with one rank below the candidate with nth rank, and

wherein n = rank of candidate (i.e. a smaller number refers to higher rank), and wherein $n = 1, 2$, and so on.

40 The higher value of the matching similarity score 504 of the pair of candidates with the largest calculated difference, D_n , in matching similarity scores is set as the first threshold 505, whereas the lower value of the matching similarity score 504 of the pair of candidates C_n and C_{n+1} with the largest calculated difference, D_n , in matching similarity scores is set as the second threshold 510.

The calculation of the difference in matching similarity scores for each pair of ranked candidates consecutively may be performed only for ranked candidates with a matching similarity score above a third threshold. For example, the third threshold may be set as a matching similarity score between 50% and 100%.

The SEER may be configured to accept the top scoring candidate C_1 as the disambiguated entity of the corresponding unresolved entity mention (corresponding frequently found entity mention), if the top scoring candidate C_1 is sufficiently higher than the second-best scoring candidate C_2 and that the top scoring candidate C_1 is one hop away from the entities in the first entity relationship map. Furthermore, although an unresolved entity mention may be disambiguated, the corresponding top scoring candidate cannot be added to the first entity relationship map if adding it would cause the expansion factor to fall outside the predetermined range.

The decision about whether an entity is 'too expansive' is dependent on the current composition of the hub entity cloud. The prospective new entity's set of one-steps is also judged for its overlap with the current one-step extended hub entity cloud: If the expansion factor is too large, for example. it would cause the one-step extended hub entity cloud to become far larger by a tuned factor, such as >20%, then this entity would not be allowed into the hub entity cloud. For example, an Audit Firm may be correctly disambiguated, but should not be allowed into the hub entity cloud because it would be linked to too many unrelated firms, thereby reducing the selectivity of the hub entity cloud based process herein adopted.

In this way, the database of known relationships adds not just a statistical measure of relatedness, but also adds a 'logical selectivity' during the process that increases precision while not reducing recall significantly. Each of the steps in the process of creating and expanding the hub entity cloud tends to have high precision (since known relationships are used to create and expand the cloud), and the cumulative effect is high recall (as the cloud expands, entities that are more distant from the Hub are added).

The expansion factor used in the decision of whether to add a disambiguated entity mention candidate to the first entity relationship map (Hub Entity Cloud) is calculated from a combination of factors stated as follows: Initial Cloud size; Expanded Cloud size; Intersection size; and Union Size. This is now elaborated with an example in Figure 3A, where a double circled "The Hour Glass Limited" is the Hub Entity 302 and single-bold-circled entities 304 and 306 are one-step entities from the Hub Entity 302 having respective relationships R1 (i.e. is subsidiary of) and R2 (i.e. is director of). Assuming the entities 302, 304 and 306 (i.e. The Hour Glass Limited, The Hour Glass (HK) Limited and Mr XYZ) are within a Current Disambiguated Hub Entity Cloud or the first entity relationship map 320, then the light solid-line circled entities 308 and 310 (i.e. Mrs XYZ and THG (HK) Limited) are one-step away from the Cloud or the first entity relationship map 320. Entity 308 (i.e. Mrs XYZ) is two steps or links from the Hub Entity

302 (via two different paths) and a dashed-line circled entity 312 (i.e. Mr PQR) is two steps from the Cloud 320.

If Entity 310 (i.e. THG (HK) Limited) and Entity 308 (i.e. Mrs XYZ) are determined as disambiguated entities, then they would likely be added to the Current Disambiguated Cloud 320. If the SEER is determining whether to add Entity 308 (i.e. Mrs XYZ) to the Current Disambiguated Cloud 320 (i.e. The Hour Glass Limited, The Hour Glass (HK) Limited and Mr XYZ), the factors: Initial Cloud size (i.e. the hub entity cloud or the first entity relationship map); Expanded Cloud size; Intersection size; and Union Size would be determined to calculate the expansion factor.

With reference to Figure 3A, the expansion factor of Mrs XYZ 308 is calculated as follows.

- The Initial Cloud size (i.e. size of 320 in Figure 3A) is 3, which comprises 3 entities, 302, 304 and 306;
- The one-step size of Mrs XYZ (including Mrs XYZ) is 3, which comprises 3 entities, 304, 306 and 308, wherein “the one-step size of Mrs XYZ (including Mrs XYZ)” is regarded as the size of a third entity relationship map (similar to the third entity relationship map of Figure 4);
- The Cloud size including the one-step size of Mrs XYZ, is 4, which comprises 4 entities, 302, 304, 306 and 308;
- The Cloud size plus its one-step expansion (the size of the Expanded Cloud or second entity relationship map 330) is 5, which comprises 5 entities 302, 304, 306, 308 and 310; and
- The Cloud size including Mrs XYZ, and their combined one-steps is also 5 (the Union value), which comprises 5 entities 302, 304, 306, 308 and 310. This union value is calculated by adding the number of entities that are present in the second entity relationship map 330 (i.e. 5, being entities 302, 304, 306, 308 and 310) and the number of entities that are present in the third entity relationship map around (and including) Mrs XYZ (i.e. 3, being entities 304, 306 and 308), and subtracting the intersection size between those two maps (i.e. 3, being entities 304, 306 and 308 that are present in both the second entity relationship map 330 and the third entity relationship map) from the sum. Therefore, in this example, the union value is $(5+3)-3 = 5$. The union value is calculated in the same manner for the other examples below.

In the present example and subsequent examples, the Expansion factor is defined as the Union divided by the size of the Expanded Cloud minus 1. In the present example, the expansion factor defines extent of expansion in size of the second entity relationship map when the third entity relationship map is combined with the second entity relationship map (i.e. the union of the second and third entity relationship maps). In another example, the expansion factor can be defined differently, such as the intersection divided over the union. Hence, depending on the application where the SEER is used, the determination of the expansion factor may vary.

Thus, the expansion factor of Mrs XYZ 308 is $(5/5 - 1) \times 100\%$, i.e. +0%. Previously, it is mentioned that if the expansion factor is too large (e.g. it would cause the one-step extended hub entity cloud to become

far larger by a tuned factor, such as >20%), then this entity would not be allowed into the hub entity cloud. In this case, the calculated expansion factor of 0% is lower than the threshold example of >20%. Therefore, Mrs XYZ 308 can be added into the Cloud 320. In this example, no lower limit of the expansion factor is set and only the upper limit of 20% is set. In another example, both lower and upper limits may be set. In summary, in the example of Figure 3A, Entity 308 (i.e. Mrs XYZ) is a good candidate for incorporating into the Cloud 320, since the 'one-steps' of Entity 308 is already completely within the Expanded Cloud. Entity 310 i.e. THG (HK) Limited is less preferred than Entity 308 as an addition to the Cloud 320 because half of its connections are external to the Cloud 320. While the size of initial set of entities is kept small in the example illustrated in Figure 3A for simplicity, it should be appreciated that the size of the initial set of entities generated by the kick-off process can be in the order of 10-20 (or more).

In view of the foregoing description, in summary, the third entity relationship map is created for each candidate (i.e. Mrs XYZ 308 in the above example), and the expanded cloud (or second entity relationship map) is also created as described above. In particular, their sizes (in terms of number of entities in each map) are determined. Thereafter, a union value is calculated from the sizes of the third and second entity relationship maps and the intersection size between the second entity relationship map and the third entity relationship map. The sizes of the union and the second entity relationship map are then used to determine the expansion factor for consideration to determine whether to add the disambiguated candidate into the existing hub entity cloud. In other words, if the expansion factor of a disambiguated candidate (e.g. Auditor) is determined to exceed an acceptable threshold (e.g. 20%), the SEER is configured to accept the disambiguated candidate, and not add it to the hub entity cloud to guard against over-expansion of the Hub Entity Cloud. This ensures that the Hub Entity Cloud remains as a useful tool for entity mention disambiguation of an input document.

With reference to Figure 4, when an entity mention is to be disambiguated as an existing entry in the database, such as entity 412, the system is configured to create a third entity relationship map 430 for the entity 412. Note that entity 412 is a candidate for disambiguating a mention in this example, and the entity 412 is one-hop away from the currently-disambiguated entities 402 and 404. The third entity relationship map 430 comprises the candidate 412 and entities 432, 402, 404 which are one hop away from the candidate 412. An expansion factor between the third entity relationship map 430 and the second entity relationship map 420 (or expanded cloud) is calculated. The expansion factor will be compared against a threshold value. In this example, the threshold is also set as 20% like the earlier example, and an entity with expansion factor lower than this threshold will be included into the first entity relationship map 310.

In determining whether to add the entity 412 into the cloud 410 illustrated in Figure 4, the following factors are determined:

- The Initial Cloud size (of 410 in Figure 4) is 3, which comprises 402, 404, and 406;

- The one-step size of Entity 412 (including Entity 412) (i.e. size of third relationship map 430) is 4, which comprises 412, 402, 404, and 432;
- The Cloud size including the one-step size of Entity 412 is 5, which includes 402, 404, 406, 412 and 432;
- The Cloud size plus its one-step expansion (i.e. size of the expanded cloud or second entity relationship map) is 7.
- The Cloud size including Entity 412, and their combined one-steps is 8 (i.e. the Union value). Specifically, the union value is calculated by adding the sum of the number of entities that are present in the second entity relationship map and the number of entities that are present in the third entity relationship map, and subtracting the intersection size between these two maps (3, being 402, 404 and 412 that are present in both the second entity relationship map and the third entity relationship map), i.e. $(7+4)-3 = 8$.

The expansion factor (based on the same definition above, i.e. (Union divided by the size of the Expanded Cloud minus 1.) is calculated as $(8/7 - 1) \times 100\% = 14.3\%$. This factor is lower than the threshold example used earlier of 20%. Therefore, the Entity 412 can be added as a Disambiguated Entity into the Hub Entity Cloud. Similarly, in this example, the expansion factor defines extent of expansion in size of the second entity relationship map when the third entity relationship map is combined with the second entity relationship map (i.e. the union of the second and third entity relationship maps).

With reference to Figure 3B, the expansion factor is calculated for a large touch-point entity. The reference numerals of elements having similar features in Fig. 3A are reused in Fig. 3B. The single-bold-circled entity 352 is a one-step entity from the double circled Hub Entity 302 having a respective relationships R7 (i.e. is auditor of). However, the entity 352 (i.e. PWC Auditor) is connected to many entities, such as entities 354, 356, 358 and 360, who are also clients (audited by) of the entity 352.

In the present example illustrated in Figure 3B,

- The Initial Cloud size (i.e. size of 320 in Figure 3B) is 3, which comprises 302, 304, and 306;
- The one-step size of PWC Auditor 352 is 6 (i.e. size of third relationship map 370), which comprises 352, 306, 360, 358, 356, and 354;
- The Cloud size including the one-step size of the PWC Auditor 352 is 8, which comprises 302, 304, 306, 352, 360, 358, 356, and 354;
- The Cloud size plus its one-step expansion (i.e. size of the Expanded Cloud or second entity relationship map) is 6, which comprises 302, 304, 306, 308, 310 and 352.
- The Cloud size including PWC Auditor, and their combined one-steps is 10 (the Union), which comprises 302, 304, 306, 352, 360, 358, 356, 354, 308 and 310. Specifically, the union value is calculated by adding the number of entities that are present in the second entity relationship map and the number of entities that are present in the third entity relationship map 370, and

subtracting their common intersection (i.e. 2, being entities 306 and 352 that are present in both the second entity relationship map and the third entity relationship map), i.e. $(6+6) - 2 = 10$.

Thus, in the example of Figure 3B, the expansion factor (using the same definition as before) is equal to $(10/6 - 1) \times 100\%$, i.e. +66%. Similarly, in this example, the expansion factor defines extent of expansion in size of the second entity relationship map when the third entity relationship map is combined with the second entity relationship map (i.e. the union of the second and third entity relationship maps).

66% is much higher than the threshold example of 20% used earlier and thus the PWC Auditor 352 has an expansion factor that is out of an acceptable range. Therefore, the entity 352 is not included in the cloud as a disambiguated entity and the resultant hub entity cloud (first entity relationship map) will be the cloud 320 shown in Figure 3A. In other words, an entity cloud or a first entity relationship map may exclude some disambiguated entities (due to the 'expansion' constraints).

Similarly, if, for instance, UOB Bank (another known large touch-point entity) also appears as an entity mention in the input document, and has a relationship with the Hub Entity, it may be determined as a Disambiguated Entity if the expansion factor of UOB Bank is not considered. After the SEER checks the expansion factor, the SEER would not incorporate the disambiguated mention 'UOB Bank' into the Hub Entity Cloud because that would cause >10% of the adult population of Singapore (estimated % of market share of the UOB Bank in Singapore) to subsequently become one-steps from the Cloud. In other words, the SEER measures the expansion factor or rate of each addition (i.e. potential disambiguated entity to be added to the Cloud), and guards against over-expansion of the Hub Entity Cloud. This ensures that the Hub Entity Cloud remains as a useful tool for entity mention disambiguation of an input document.

Other than the edge-of-cloud disambiguation, the SEER may also be configured to use the content of the current document (e.g. using the sentences surrounding a particular entity mention) to form a 'word cloud fingerprint' around each entity under consideration. The SEER may also be configured to use the already-disambiguated entities in the document and rank the choices of entity for disambiguation based on the frequency that other individual entities are commonly associated with each disambiguated entity in the list of currently-disambiguated entities. These features also help to ensure that the hub entity cloud is updated with useful disambiguated entities.

Returning to Figure 1, the above-mentioned processing steps S114 and S116 iteratively examines NER mentions from the list of unresolved mentions, and determines whether/how they can be added to the cloud (first entity relationship map) obtained for the document. After each successful addition of a disambiguated entity to the hub entity cloud, the remaining unresolved mentions would be re-prioritised. This is advantageous because every determination made about which disambiguated entities are present in the document affects subsequent decisions about which other named entities

(candidates for unresolved mentions) may be present in the document. Effectively, during re-prioritisation, the system is configured to re-assess the list of unknown/unmatched NER given the current state of matching that has been done (i.e. by making reference to the list of currently disambiguated entities).

4) Proactive searching

Proactive searching for additional entity mentions in the input document is initiated by the system (e.g. 200 in Figure 2) at a step S118 after the disambiguation process of matching a mention to an entity (i.e. mention -> entity) in steps 114 and 116 is exhausted.

The step S118 focuses on the current (updated) expanded hub entity cloud (i.e. the second entity relationship map), which is the version of the expanded cloud after steps 114 and 116 are completed. At this point in the process, many entities in the expanded cloud (second entity relationship map) may not have mentions associated with them, in particular, the entities that are one step from the cloud (first entity relationship map). However, it is possible for the names of the entities in the expanded cloud to appear in the text of the input document, but have not have been identified by the NER process. This is commonly due to formatting or incomplete sentences within the document, or that the entities in the expanded cloud contain unusual names. Therefore, in order to increase recall, fuzzy full-text searches are performed against the document again. In one example, the fuzzy search is performed based on one-step relationships to the updated entity cloud (i.e. updated first entity relationship map) to surface additional mentions, which can then be disambiguated using the disclosed techniques in the present disclosure with the knowledge of which entities are associated with the mention under consideration. The step S118 prioritizes recall, with the precision being derived from the limited number of 'known valid' entity names being searched. The step S118 takes place before the 'co-referencing step S120' in the flow chart of Figure 1.

Turning back to Figure 3A, if "Mrs PQR" (wife of disambiguated entity 314, i.e. Mr PQR) is a mention in the input document, the system would be able to reliably identify her as a Disambiguated Entity after processing a mention 312 in the input document that has been successfully disambiguated as Mr PQR (e.g. 314 in Figure 3A). Furthermore, if Mr PQR and Mrs PQR have a son "Master PQR", the system would search parsed text of the input document for the string "Master PQR" during the "Proactively Searching for Entities" step S118. This search for "Master PQR" would be performed even if there are no NER matching algorithm that can pick up "Master PQR" as a mention.

While fuzzy matching may be employed in the state of the art to allow for flexibility in matching entities, the step S118 is configured to find matches actively based on a database entity that may be in the document (due to the proximity of relationship to the currently-disambiguated entities). In other words, the SEER is looking for text strings in the document based on relationships and entity names in the database, instead of searching for mentions in the database based on strings found in the text. This is desirable because the fuzzy search is performed on possible names in the document that none of

the NER matching algorithms identified as mentions at the step S108. Note that all non-disambiguated entity strings are cast aside at the step S118.

Step S118 is implemented in the disclosed entity disambiguation process as a final 'mop up' to actively seek entity mentions that are driven by the verified entity relationship data in the database and the list of currently disambiguated entities. For instance, it may be that in the text, due to poor layout, a name "Singapore Dragon Boat Association" is laid out across columns, so that the only NER found are "Singapore Dragon" and "Boat Association". Neither of these terms are sufficient to disambiguate to the true entity on their own. However, knowing that the CFO of a disambiguated company is a member of the Singapore Dragon Boat Association (and using the links within the database of relations between already disambiguated entities), the SEER can proactively look for the string "Singapore Dragon Boat Association" in the document. Given the relationship that is known to exist, it is unlikely that "Singapore Dragon Boat Association" would appear by chance in the document, leading this method (i.e. in particular, step S118) to have high precision and have increased recall for 'difficult to find' entities. This is in contrast to a matching performed in a forward direction ("Singapore Dragon" or "Boat Association") across the database. Note that matching in a forward direction would be far more error prone.

An example of the algorithm used in step S118 is set out below.

For each disambiguated entity that is found by the entity disambiguation method from S102 to S116, a list of "one-step" entities is generated using the verified entity relationship data in the database. 'One-step' entities are entities that are one hop away from said disambiguated entity, and can be obtained from the current second entity relationship map (i.e. Expanded Cloud) available just prior to S118.

Thereafter, for each alias of each of these related 'one-step' entities, the SEER is configured to search the document for string matches that are similar to said alias of said 'one-step' entity with latitude allowed (e.g. a company name split across two columns of text) for character-wise errors. If there is a match, the string mention can be disambiguated as the corresponding validated entity in the second entity relationship map immediately. The disambiguated entity mention can also be added into the first entity relationship map immediately and the second entity relationship map is updated accordingly, because the mention is expected to be appear in the input document.

This means that fuzzy full text search is performed against the input document using each disambiguated entity mention and each of the one or more entities that is one hop away from the disambiguated entity mention in the second entity relationship map, wherein a fuzzy matched string in the document found via the fuzzy full text search with an acceptable matching similarity score that is not listed as a disambiguated entity mention in the second entity relationship map is updated as a disambiguated entity mention. The acceptable matching similarity score can be, for example, a matching similarity score within between 50% to 100% or any narrower range between 50% to 100%.

Subsequently, after step S118, the disambiguation process moves on to Entity Co-Referencing at a step S120. At step S120, the SEER is configured to identify short forms of names (such as “Mr Tan Ah Kow” when he is subsequently referred to as “Mr Tan” in the said document). In addition, the SEER has to link the words “he”, “she”, “it”, “we” or “they” correctly to the said disambiguated entity within the document. By tracking these co-references, the SEER is able to detect/form relationships at step S124 with those referenced entities when they are mentioned in different sentences.

Optionally, if the SEER detects table(s) in the document, Table Analysis is initiated at S122. Conventional methods of table detection can be used.

After table analysis at step S122, Relationship Analysis is conducted at a step S124 to determine potential relationship mentions (e.g. R1 to R6 in Figure 3A) between pairs of entities within sentences in the document. Once disambiguated entities have been identified within each sentence (whether from the parsed text of the document, or derived from tabular data), the data is inputted into a Relationship Analysis algorithm which attempts to detect, classify and disambiguate relationships within the document text. To detect relationships from textual descriptions and classify them according to their nature in the SEER, text relationship classification models have been trained to recognise approximately 100 different directional relationships (such as “Is Director Of”, “Is Child Of”, etc). The training data consists of both “ideal sentences” and “real world sentences” captured from previous documents that the SEER has analysed, annotated by humans, including examples corrected by humans through a human-in-the-loop UI interface. One example of an idealized sentence is “John Smith is a director of ABC Corp.”, which is also an idealized example of a “Is Director Of” relationship). One example of a real world sentence is “Mr. David Ho Hing Yuen was appointed to the Board of Directors of DBS Bank (Hong Kong) Limited on 20 March 2019.”

The relationship model covers all aspects of what may be considered a real world tangible relationship category, including Family, Ownership, Board, Management, Commercial, Professional, Transaction and Education. The list of Relationship Categories may be edited or added to over time.

Once a text relationship has been detected and classified, it is disambiguated against validated links (or relationships) between different validated entities in the verified entity relationship data of the database. Disambiguation is performed by matching the fingerprint of a relationship (entities, relationship category) against those in the knowledge database. The SEER may also be configured to regularize graphs of entities and relationships derived from the input document according to a set of rules. For instance, a Director who is also an Employee of the Company cannot be an Independent Director. The output of the Relationship Analysis process in S124 are disambiguated relationships backed by textual evidence, which will be inputted to an output stage at a step S126. The Relationship Analysis step S124 may be subject to iterative refinement to continuously update the relationship data.

Disambiguated entities and disambiguated relationships between the disambiguated entities are recorded into the database at the step S126. In one example, information associated with the entities in the database that corresponds to the disambiguated entities may be updated. For instance, the information of disambiguated entity is updated to include a comment “mentioned in a document labelled as Annual Report 2013, of The Hour Glass Limited”, and the disambiguated relationship between two disambiguated entities, such as “a-subsiary-of”, may also be updated as “last validated in the document labelled as Annual Report Annual Report 2013, of The Hour Glass Limited”. In another example, if the document shows evidence that the disambiguated relationship between two disambiguated entities has changed from the validated relationship in the database, the system may be configured to change the information accordingly and flag the change for further curation. For instance, “chairman-of” may be changed to “chairman-of” with a defined end date. A line representing the validated relationship between the validated entities may also be formatted to indicate the change, such as from a solid line to a dotted line.

The step S126 may be incorporated with an optional human-curation step. The results of the analysis are presented together with the current verified entity relationship data of the database, for further curation by a human analyst and grants approval for incorporation of the changes.

To facilitate this human-curation step, a user interface is provided by the system so that the source sentences and outputs of the relevant models obtained from the document are presented to the human analyst for interpretation and/or validation. The user interface allows the human analyst to search for and access processed documents. When accessing a document, the user is presented with all disambiguated entities and disambiguated relationships and the supporting evidence extracted from the document text. This is particularly useful for the validation of relationships, because the human analyst (user) can view all validated relationships of the current verified entity relationship data from the knowledge database that can possibly match a predicted relationship text. This enables the user to quickly identify relationships that are not positively predicted by the system and correct it if necessary.

The user interface also enables human analyst (or user) to directly correct each model's decisions, and the corrected data is then added to the pool of training examples for subsequent model training. The verified entity relationship map preferably contains a sufficient number of validated entities and validated relationships between validated entries. Visualisation report indicating a relationship map retrieved from the document and a corresponding relationship map of the validated entities from the knowledge database can also be generated for comparison with for KYC process of an organization.

The database comprising the verified entity relationship map can be used in the disambiguation of entities in a company report. This database (knowledge base) can contain many entities that appear very similar on the surface (e.g. Apple Computer Inc. vs Apple (USA) Corp.), but are vital to distinguish for KYC purposes. The database also aims to cover a comprehensive selection of financially active people within a given country, rather than just notable people (such as those that might be featured in a news article). For instance, there are listed companies where in excess of 50% of the board of

directors are named “Mr Tan” (being an extremely common name in Singapore) - and each of these separate Mr Tan entities needs to be uniquely and separately identified.

The present disclosure is particularly useful to disambiguate entity mentions in a lengthy document about a single subject as it uses many different techniques (e.g. steps S102 to S126 in Figure 1) to disambiguate entity mentions. For instance, a company annual report, which can be about a few hundred pages, would require these different techniques to perform entity disambiguation to ensure accuracy. Although the present disclosure is particularly useful to disambiguate entity mentions in a lengthy document about a single subject, of course, the present disclosure can also be used to disambiguate entity mentions in short news stories or a collection of webpages such as Wikipedia articles.

A client device or apparatus through which a user can launch the SEER application to inspect, correct and/or approve one or more documents, a server machine that runs the services associated with the software components of the SEER application, and a computing device used by a human analyst are examples of a user device or apparatus that interacts with the SEER application. The user device or apparatus can be for instance, a desktop computer or a mobile device such as a mobile smartphone, laptop, notebook, and the like.

Estimated statistical results of the performance of the SEER is presented below. Each innovation described above i.e. 1) Kick-off Matching, 2) Hub Entity, 3) One-Step Analysis in Edge-of-cloud disambiguation, and 4) Proactive searching is introduced in turn below. An estimate of the average effects on the F1 scores for disambiguated entity for each of the different innovations is as follows:

Kick-off Matching:	+7.75%
Hub Entity:	+0.23%
One-Step Analysis in Edge-of-cloud disambiguation:	+7.24%
Proactive searching:	+1.24%

The F1 Score is based on the formula, $2 * [(precision * recall) / (precision + recall)]$. It is also called the F Score or the F Measure. Generally, the F1 score conveys the balance between the precision and the recall.

Figure 7 shows in more detail an example of a processor in the user device (client device, device used by analyst, or apparatus) or a processor in a server (or apparatus) that is configured to run an algorithm to extract entity relationships (known as the SEER application or the SEER in the present disclosure). The SEER comprises a number of interconnected software components, with the flow of data between them orchestrated by a processor that is configured to extract structured data about people, companies and events from unstructured text input. The SEER may have hardware components as well.

The processor may comprise a processing unit 702 for processing software including one or more computer programs for running one or more computer/server applications to enable a backend logic flow or the method or methods for carrying out the steps as disclosed above with reference to the earlier Figures.

5

Furthermore, the processing unit 702 may include user input modules such as a computer mouse 736, keyboard/keypad 704, and/or a plurality of output devices such as a display device 708. The display of the display device 708 may be a touch screen capable of receiving user input as well.

10 The processing unit 702 may be connected to a computer network 712 via a suitable transceiver device 714 (i.e. a network interface), to enable access to e.g. the Internet or other network systems such as a wired Local Area Network (LAN) or Wide Area Network (WAN). The processing unit 702 may also be connected to one or more external wireless communication enabled devices 734 via a suitable wireless transceiver device 732 e.g. a WiFi transceiver, Bluetooth module, Mobile telecommunication transceiver
15 suitable for Global System for Mobile Communication (GSM), 3G, 3.5G, 4G telecommunication systems, and the like. Through the computer network 712, the processing unit 702 can gain access to one or more storages i.e. data storages, databases, data servers and the like connectable to the computer network 712 to retrieve and/or store data in the one or more storages.

20 The processing unit 702 may include a processor 718, a Random Access Memory (RAM) 720 and a Read Only Memory (ROM) 722. The processing unit 702 may also include a number of Input/Output (I/O) interfaces, for example I/O interface 738 to the computer mouse 736, a memory card slot 716, I/O interface 724 to the display device 708, and I/O interface 726 to the keyboard/keypad 704.

25 The components of the processing unit 702 typically communicate via an interconnected bus 728 and in a manner known to the person skilled in the relevant art.

The computer programs may be supplied to the user of the processing unit 702, or the processor (not shown) of one of the one or more external wireless communication enabled devices 734, encoded on
30 a data storage medium such as a CD-ROM, on a flash memory carrier or a Hard Disk Drive, and are to be read using a corresponding data storage medium drive of a data storage device 730. Such computer or application programs may also be downloaded from the computer network 712. The application programs are read and controlled in its execution by the processor 718. Intermediate storage of program data may be accomplished using RAM 720.

35

In more detail, one or more of the computer or application programs may be stored on any non-transitory machine- or computer- readable medium. The machine- or computer- readable medium may include storage devices such as magnetic or optical disks, memory chips, or other storage devices suitable for interfacing with a general purpose computer. The machine- or computer- readable medium may also
40 include a hard-wired medium such as that exemplified in the Internet system, or wireless medium such

as that exemplified in the Wireless LAN (WLAN) system and the like. The computer program when loaded and executed on such a general-purpose computer effectively results in an apparatus that implements the steps of the computing methods in examples herein described.

5 In summary, examples of the present disclosure may include the following features.

A method (e.g. S100 in Figure 1) for disambiguating a plurality of entity mentions in a document, the method comprising the following steps:

- 10 (a) parsing text (e.g. 202 in Figure 2) of a document to obtain a plurality of entity mentions (e.g. 204a, 204b in Figure 2);
- (b) determining one or more matched entity mention by matching each of the plurality of entity mentions with each entity listed in a database (e.g. S110 in Figure 1), wherein the database (e.g. 214 in Figure 2) comprises verified entity relationship data indicating links between different entities;
- 15 (c) determining a score for each matched entity mention (e.g. 210a, 210b in Figure 2, 402, 404 and 406 in Figure 4) based on entity type and number of links the matched entity mention has with other matched entity mentions; and
- (d) determining the matched entity mention with highest score as a hub entity (e.g. S112 in Figure 1), wherein the hub entity (e.g. 302 in Figures 3A and 3B, 402 in Figure 4) is deemed as a disambiguated entity mention.

20

The score for each matched entity mention (e.g. 210a, 210b in Figure 2, 402, 404 and 406 in Figure 4) may be determined based on entity type and number of links the matched entity mention has with other matched entity mentions by:

- 25 (e) determining, for each matched entity mention, one or more entities that are indicated as a first predetermined number of hops away from the corresponding matched entity in the verified entity relationship data;
- (f) obtaining a weight value (e.g. w_1 , w_2 , w_3 , w_4 , w_5 and w_6 in Figure 4) for each of the one or more entities, wherein the weight value was predetermined based on type of entity;
- 30 (g) if there are more than one entities determined for the matched entity mention, calculating for the matched entity mention, a weighted sum of the weight values of the more than one entities;
- (h) if there are more than one matched entity mentions, identifying the matched entity mention that has the calculated weighted sum satisfying a first threshold as a hub entity candidate;
- (i) determining a score for each hub entity candidate by counting number of links between said hub entity candidate and each matched entity mention found in the verified entity relationship data, wherein the links to be counted are the links with a second predetermined number of hops between said hub entity candidate and each matched entity mention; and
- 35 (j) identifying the hub entity candidate with highest score as the hub entity (e.g. 302 in Figures 3A and 3B, 402 in Figure 4).

The method may further comprise:

- (a) identifying any matched entity mention, which is a large entity (e.g. 352 in Figure 3B) with number of links above a first threshold; and
- (b) creating a first entity relationship map (e.g. 320 in Figures 3A and 3B, 410 in Figure 4) for the document using the hub entity and the one or more matched entity mention (e.g. 304, 306 in Figures 3A and 3B, 404 and 406 in Figure 4) linked to the hub entity that is not the large entity.

The method may further comprise creating a second entity relationship map (e.g. 330 in Figure 3A, 420 in Figure 4) using the entities in the first entity relationship map, and entities that are one hop away from the entities in the first entity relationship map.

The method may further comprise:

- (a) obtaining a predetermined number of most frequently found entity mentions from the parsed text of the document and fuzzy match each frequently found entity mentions against each entity listed in the database to return a list of disambiguated entity mention candidates (e.g. $C_1, C_2, C_n, C_{n+1}, C_{n+2}$ in Figure 5);
- (b) determining top candidates (e.g. C_1, C_2, C_n in Figure 5) with matching similarity scores from the list of disambiguated entity mention candidates that are above or equal to a second threshold (e.g. 505 in Figure 5) ;
- (c) creating, for each top candidate, a third entity relationship map (e.g. 370 in Figure 3B, 430 in Figure 4) using said top candidate and one or more entities one hop away from said top candidate;
- (d) determining, for each top candidate, an expansion factor defining extent of expansion in size of the second entity relationship map when the third entity relationship map is combined with the second entity relationship map; and
- (e) determining the top candidate with expansion factor within an acceptable predetermined range of values as a disambiguated entity mention.

The method may further comprise:

updating the first entity relationship map by adding the disambiguated entity mention, which is the top candidate with expansion factor within the acceptable predetermined range of values, to the first entity relationship map; and

updating the second entity relationship map by adding one or more entity that is one hop away from the disambiguated entity mention in the verified entity relationship data, which is the top candidate with expansion factor within the acceptable predetermined range of values, to the second entity relationship map.

The second threshold (e.g. 505 in Figure 5) may be calculated by:

ranking all the disambiguated entity mention candidates (e.g. $C_1, C_2, C_n, C_{n+1}, C_{n+2}$ in Figure 5) according to matching similarity scores;

calculating a difference in matching similarity scores for each pair of ranked candidates consecutively, according to a formula, $D_n = C_n - C_{n+1}$, wherein

D_n = difference between candidate with n th rank and candidate with one rank below the n th rank,
 C_n = candidate with n th rank,

5 C_{n+1} = candidate with one rank below the candidate with n th rank

$n = 1, 2$, and so on, wherein n = rank of candidate; and

setting, as the second threshold, the higher value of the matching similarity score of the pair of candidates with the largest calculated difference in matching similarity scores.

10 Calculating of the difference in matching similarity scores for each pair of ranked candidates consecutively may be performed only for ranked candidates with a matching similarity score above a third threshold (e.g. 510 in Figure 5).

The first predetermined number of hops may be one hop.

15

The second predetermined number of hops may be one hop.

The weight value may be predetermined based on the types of entities as follows:

Employee of an entity;

20 Executive director of the entity;

Shareholder of the entity; and

Subsidiary entity of the entity;

where the Executive director has the greatest weight value, the Subsidiary entity has the least weight value, and the Employee and Shareholder have the same weight value or weight values

25 between the greatest weight value and the least weight value.

Optionally, the types of entities may include parent entity or holding entity of the entity, and the parent entity or the holding entity may be assigned a negative weight value.

30 A fuzzy full text search may be performed against the document using each disambiguated entity mention and each of the one or more entities that is one hop away from the disambiguated entity mention in the updated second entity relationship map, wherein a fuzzy matched string in the document found via the fuzzy full text search with a predetermined acceptable matching similarity score that is not listed as a disambiguated entity mention in the updated second entity relationship map is updated as a
35 disambiguated entity mention.

There are various ways of formulating the expansion factor, and may depend on user-defined conditions set to address different applications.

The expansion factor of each top candidate may be calculated based on a size of the first entity relationship map, a size of the second entity relationship map, a size of the third entity relationship map, an intersection size and/or a union size. In one example, the expansion factor is determined based on a value calculated by dividing the intersection size divided over the union size. In this example, the intersection size is calculated by adding number of entities in the first entity relationship map and the number of entities in the third entity relationship map to obtain a value, and subtracting said value by number of entities in a cloud formed by combining the first entity relationship map and the third entity relationship map. Subsequently, the union value is calculated by (i) subtracting said intersection size from the sum of the number of entities present in the second entity relationship map and the number of entities that are present in the third entity relationship map, and subtracting said value by 1.

In another example, the expansion factor is defined as the Union divided by the size of the second entity relationship map (or Expanded cloud) minus 1,

wherein the size of the third entity relationship map is determined by the number of disambiguated entities in the third entity relationship map,

wherein the size of the second entity relationship map is determined by the number of entities in the second entity relationship map,

wherein the intersection size is determined by the number of entities that are present in both the second entity relationship map and the third entity relationship map,

wherein the union value is obtained by:

adding the number of entities that are present in the second entity relationship map and the number of entities that are present in the third entity relationship map, and

subtracting the number of entities that are present in both the second entity relationship map and the third entity relationship map.

The method may comprise a step of categorizing the plurality of entity mentions to filter entity mentions that are categorized as junk mention, wherein a junk mention is a predetermined invalid entity deemed to be not a corporate entity or an individual entity.

An apparatus (e.g. 702 in Figure 7) for disambiguating a plurality of entity mentions in a document, wherein the apparatus comprises:

a processor (e.g. 718 in Figure 7) for executing instructions in a memory (720 in Figure 7) to operate the apparatus to:

(a) parse text (e.g. 202 in Figure 2) of a document to obtain a plurality of entity mentions (e.g. 204a, 204b in Figure 2);

(b) determine one or more matched entity mention by matching each of the plurality of entity mentions with each entity (e.g. 210a, 210b in Figure 2) listed in a database, wherein the database (e.g. 214 in Figure 2) comprises verified entity relationship data indicating links between different entities;

(c) determine a score for each matched entity mention (e.g. 210a, 210b in Figure 2, .402, 404 and 406 in Figure 4) based on entity type and number of links the matched entity mention has with other matched entity mentions; and

(d) determine the matched entity mention with highest score as a hub entity (e.g. 302 in Figures 3A and 3B, 402 in Figure 4), wherein the hub entity is deemed as a disambiguated entity mention.

A system for disambiguating a plurality of entity mentions in a document, wherein the system comprises the aforementioned apparatus. The system may comprise a server configured to store verified entity relationship data indicating links between different verified entities and a display device for displaying validated relationships of the verified entity relationship data, the entities disambiguated by the system, disambiguated relationships between the disambiguated entities and the supporting evidence extracted from the document.

The database (knowledge base) may contain many very similar entities, each starting with a common corporate identification, such as Apple Computer, Apple Inc, Apple Holdings (Malaysia) Snd Bhd, Apple N.A. Property REIT 15-2 LLC, etc. Furthermore, the database may also cover a far more comprehensive selection of people within a country, rather than just notable people that might be featured in a news article.

The disclosed method performed by the SEER in the present disclosure may be used to disambiguate entity mentions in a lengthy document about a single subject, such as a company report.

In the specification and claims, unless the context clearly indicates otherwise, the term “comprising” has the non-exclusive meaning of the word, in the sense of “including at least” rather than the exclusive meaning in the sense of “consisting only of”. The same applies with corresponding grammatical changes to other forms of the word such as “comprise”, “comprises” and so on.

While the invention has been described in the present disclosure in connection with a number of embodiments and implementations, the invention is not so limited but covers various obvious modifications and equivalent arrangements, which fall within the purview of the appended claims. Although features of the invention are expressed in certain combinations among the claims, it is contemplated that these features can be arranged in any combination and order.

Claims

1. A method for disambiguating a plurality of entity mentions in a document, the method comprising the following steps:

- (a) parsing text of a document to obtain a plurality of entity mentions;
- (b) determining one or more matched entity mention by matching each of the plurality of entity mentions with each entity listed in a database, wherein the database comprises verified entity relationship data indicating links between different entities, and the entity mention of the plurality of entity mentions having exact match with an entity listed in the database is determined as a matched entity mention;
- (c) determining a score for each matched entity mention based on entity type and number of links the matched entity mention has with other matched entity mentions; and
- (d) determining the matched entity mention with highest score as a hub entity, wherein the hub entity is deemed as a disambiguated entity mention,

wherein the score for each matched entity mention is determined based on entity type and number of links the matched entity mention has with other matched entity mentions by:

- (e) determining, for each matched entity mention, one or more entities that are indicated as a first predetermined number of hops away from the corresponding matched entity in the verified entity relationship data;
- (f) obtaining a weight value for each of the one or more entities, wherein the weight value was predetermined based on type of entity;
- (g) if there are more than one entities determined for the matched entity mention, calculating for the matched entity mention, a weighted sum of the weight values of the more than one entities;
- (h) if there are more than one matched entity mentions, identifying the matched entity mention that has the calculated weighted sum satisfying a first threshold as a hub entity candidate;
- (i) determining a score for each hub entity candidate by
 - counting number of links between said hub entity candidate and each matched entity mention found in the verified entity relationship data, wherein the links to be counted are the links with a second predetermined number of hops between said hub entity candidate and each matched entity mention, or
 - making reference to the verified entity relationship data to count number of entities connected to each hub entity candidate within a second predetermined number of hops; and
- (j) identifying the hub entity candidate with highest score as the hub entity.

2. The method of claim 1, wherein the method further comprises:

- (a) identifying whether any matched entity mention is a large entity, wherein a matched entity mention is identified as a large entity if the number of links the matched entity has with other matched entity mentions is above or equal to a first threshold; and
- (b) creating a first entity relationship map for the document using the hub entity and the one or more matched entity mention linked to the hub entity that is not the large entity.

3. The method of claim 2, wherein the method further comprises creating a second entity relationship map using the entities in the first entity relationship map, and entities that are one hop away from the entities in the first entity relationship map.

4. The method of claim 3, wherein the method further comprises:

- (a) obtaining a predetermined number of most frequently found entity mentions from the parsed text of the document and fuzzy match each frequently found entity mentions against each entity listed in the database to return a list of disambiguated entity mention candidates;
- (b) determining top candidates with matching similarity scores from the list of disambiguated entity mention candidates that are above or equal to a second threshold;
- (c) creating, for each top candidate, a third entity relationship map using said top candidate and one or more entities one hop away from said top candidate;
- (d) determining, for each top candidate, an expansion factor defining the extent of expansion in size of the second entity relationship map when the third entity relationship map is combined with the second entity relationship map; and
- (e) determining the top candidate with expansion factor within an acceptable predetermined range of values as a disambiguated entity mention.

5. The method of claim 4, wherein the method further comprises:

updating the first entity relationship map by adding the disambiguated entity mention, which is the top candidate with expansion factor within the acceptable predetermined range of values, to the first entity relationship map; and

updating the second entity relationship map by adding one or more entity that is one hop away from the disambiguated entity mention in the verified entity relationship data, which is the top candidate with expansion factor within the acceptable predetermined range of values, to the second entity relationship map.

6. The method of any one of claim 4 or 5, wherein the second threshold is calculated by:

ranking all the disambiguated entity mention candidates according to matching similarity scores;

calculating a difference in matching similarity scores for each pair of ranked candidates consecutively, according to a formula, $D_n = C_n - C_{n+1}$, wherein

D_n = difference between candidate with n th rank and candidate with one rank below the n th rank,

C_n = candidate with n th rank,

C_{n+1} = candidate with one rank below the candidate with n th rank

$n = 1, 2$, and so on, wherein n = rank of candidate; and

setting, as the second threshold, the higher value of the matching similarity score of the pair of candidates with the largest calculated difference in matching similarity scores.

7. The method according to claim 6, wherein calculating of the difference in matching similarity scores for each pair of ranked candidates consecutively is performed only for ranked candidates with a matching similarity score above or equal to a third threshold.

8. The method according to any one of claims 1 to 7, wherein the first predetermined number of hops is one hop.

9. The method according to any one of claims 1 to 8, wherein the second predetermined number of hops is one hop.

10. The method of any one of claims 1 to 9, wherein the weight value was predetermined based on the types of entities as follows:

Employee of an entity;

Executive director of the entity;

Shareholder of the entity; and

Subsidiary entity of the entity;

where the Executive director has the greatest weight value, the Subsidiary entity has the least weight value, and the Employee and Shareholder have the same weight value or weight values between the greatest weight value and the least weight value.

11. The method of any one of claims 5 to 10, wherein a fuzzy full text search is performed against the document using each disambiguated entity mention and each of the one or more entities that is one hop away from the disambiguated entity mention in the updated second entity relationship map, wherein a fuzzy matched string in the document found via the fuzzy full text search with a predetermined acceptable matching similarity score that is not listed as a disambiguated entity mention in the updated second entity relationship map is updated as a disambiguated entity mention.

12. The method of any one of the claims 4 to 11, wherein the expansion factor of each top candidate is calculated based on a size of the third entity relationship map, a size of the second entity relationship map, an intersection size and a union size,

wherein the size of the third entity relationship map is determined by the number of disambiguated entities in the third entity relationship map,

wherein the size of the second entity relationship map is determined by the number of entities in the second entity relationship map,

wherein the intersection size is determined by the number of entities that are present in both the second entity relationship map and the third entity relationship map,

wherein the union size is obtained by:

adding the number of entities that are present in the second entity relationship map and the number of entities that are present in the third entity relationship map, and subtracting

the number of entities that are present in both the second entity relationship map and the third entity relationship map.

13. The method of any one of the preceding claims, wherein the method comprising a step of categorizing the plurality of entity mentions to filter entity mentions that are categorized as junk mention, wherein a junk mention is a predetermined invalid entity deemed to be not a corporate entity or an individual entity.

14. An apparatus for disambiguating a plurality of entity mentions in a document, wherein the apparatus comprises:

a processor for executing instructions in a memory to operate the apparatus to:

- (a) parse text of a document to obtain a plurality of entity mentions;
- (b) determine one or more matched entity mention by matching each of the plurality of entity mentions with each entity listed in a database, wherein the database comprises verified entity relationship data indicating links between different entities, and the entity mention of the plurality of entity mentions having an exact match with an entity listed in the database is determined as a matched entity mention;
- (c) determine a score for each matched entity mention based on entity type and number of links the matched entity mention has with other matched entity mentions; and
- (d) determine the matched entity mention with highest score as a hub entity, wherein the hub entity is deemed as a disambiguated entity mention, wherein the apparatus is operable to:
 - (e) determine, for each matched entity mention, one or more entities that are indicated as a first predetermined number of hops away from the corresponding matched entity in the verified entity relationship data;
 - (f) obtain a weight value for each of the one or more entities, wherein the weight value was predetermined based on type of entity;
 - (g) if there are more than one entities determined for the matched entity mention, calculate for the matched entity mention, a weighted sum of the weight values of the more than one entities;
 - (h) if there are more than one matched entity mentions, identify the matched entity mention that has the calculated weighted sum satisfying a first threshold as a hub entity candidate;
 - (i) determine a score for each hub entity candidate by
 - counting number of links between said hub entity candidate and each matched entity mention found in the verified entity relationship data, wherein the links to be counted are the links with a second predetermined number of hops between said hub entity candidate and each matched entity mention, or
 - making reference to the verified entity relationship data to count number of entities connected to each hub entity candidate within a second predetermined number of hops; and
 - (j) identify the hub entity candidate with highest score as the hub entity.

15. A system for disambiguating a plurality of entity mentions in a document, wherein the system comprises the apparatus as claimed in claim 14.